Case Study 2

Auto-scaling & Cost Optimization

Scenario

“A media-streaming client experienced unpredictable traffic surges during live events. At peak times, the system slowed down; off-peak, the client was overpaying for idle infrastructure.”

- Compute costs spiked during events.
- Underutilized VMs in off-hours.
- Storage costs growing due to old data.

Challenges

- Performance risk → Application lagged during peak demand.
- High cost → Paying for unused compute/storage at off-peak.
- Inefficient resource management → No clear policy for scaling or data archiving.

Proposed Solution

"We recommended an elastic infrastructure model to balance cost with performance."

- Introduced Kubernetes HPA (Horizontal Pod Autoscaler) to scale on demand.
- Used async job/event depth metrics to pre-warm pods and avoid cold starts.
- Segregated storage: frequently accessed → hot tier, old data → archive storage (cheaper).

Expected Business Outcomes

- Seamless customer experience during traffic surges.
- 30–40% cost reduction by scaling down off-peak + archiving data.
- Operational efficiency → infra automatically adapts to load.